Discrete Applied Mathematics 46 (1993) 275–281
North-Holland

Note

A characterization of overlap-free morphisms*

J. Berstel

LITP Institut Blaise Pascal, Université Pierre et Marie Curie, 4, place Jussieu, F-75252 Paris Cedex 05, France

P. Séébold

Faculté de Mathématiques et Informatique, LAMIFA, 33, rue Saint Leu, F-80039 Amiens Cedex, France

Received 3 November 1992
Revised 21 May 1993

Abstract

We prove that a morphism h over a two-letter alphabet $\{a, b\}$ is overlap-free, i.e., maps overlap-free words into overlap-free words, iff the word $h(abbabaab)$ is overlap-free. As a consequence, we obtain a simple proof of the fact that the only infinite overlap-free words that can be obtained by iterating a morphism are the Thue–Morse sequence and its opposite.

Résumé

Nous prouvons qu'un morphisme h sur un alphabet à deux lettres $\{a, b\}$ est sans chevauchement, c'est-à-dire envoie les mots sans chevauchement sur des mots sans chevauchement si et seulement si le mot $h(abbabaab)$ est sans chevauchement. Comme conséquence, nous obtenons une preuve simple du fait que les seuls mots infinis sans chevauchement qui peuvent être obtenus par itération d'un morphisme sont le mot de Thue–Morse et son opposé.

1. Introduction

Thue was the first to show, in 1912 [15] the existence of an infinite overlap-free word over two letters. He indeed proved that the infinite word

$$t = abbabaabbaababbab \ldots$$

Correspondence to: Professor J. Berstel, LITP Institute Blaise Pascal, Université Paris 6, 4 Place Jussieu, F- 75252 Paris Cedex 05, France.

* Partially supported by the PRC "Mathématique et Informatique".

that is now called the Thue–Morse sequence, is overlap-free. This was rediscovered later by Morse [9], Morse and Hedlund [10], Arson [1] and others. A fairly complete description of all infinite overlap-free words over two letters is given in Fife [4]. Recent results can be found in [2, 7, 11, 14]. Surveys are in [8, 12]. Thue's proof consists in showing that a special morphism μ over two letters, defined by $\mu(a) = ab$, $\mu(b) = ba$, is overlap-free: for any overlap-free word x, the word $\mu(x)$ is overlap-free. More generally, let h be an overlap-free morphism over a two-letter alphabet $\{a, b\}$, i.e., a morphism that maps overlap-free words into overlap-free words. Thue proved that h is basically a power of the morphism μ.

The aim of this note is to show that, in order to test whether a morphism h is overlap-free, it suffices to check that the single word $h(abbabaab)$ is overlap-free. Thue's characterization is an immediate consequence of this result. Moreover, we obtain a simple proof of a stronger result, first proved by Séébold [13], namely that the Thue–Morse sequence t and its opposite, obtained by exchanging a and b, are the only infinite overlap-free words that can be generated by iterating a morphism, or equivalently, are the only infinite overlap-free words that are fixed points of a non-trivial overlap-free morphism.

2. Definitions

Let $A = \{a, b\}$ be a two-letter alphabet. The empty word is denoted by ε, the length of a word u is denoted by $|u|$. A *morphism* is a mapping h from A^* into itself such that $h(uv) = h(u)h(v)$ for all words u, v. A morphism is *nonerasing* if neither $h(a)$ nor $h(b)$ is the empty word. In the sequel, all morphisms will be supposed to be distinct from the null morphism which maps all letters into the empty word. Consider the morphism μ from the free monoid A^* into itself defined by

$$\mu(a) = ab, \qquad \mu(b) = ba.$$

Setting, for $n \geq 0$,

$$u_n = \mu^n(a), \qquad v_n = \mu^n(b)$$

one gets

$$u_0 = a \qquad v_0 = b$$

$$u_1 = ab \qquad v_1 = ba$$

$$u_2 = abba \qquad v_2 = baab$$

$$u_3 = abbabaab \qquad v_3 = baababba$$

and more generally $u_{n+1} = u_n v_n$, $v_{n+1} = v_n u_n$ and $u_n = E(v_n)$, $v_n = E(u_n)$, where E is the morphism that exchanges a and b. The word $E(w)$ is called the *opposite* of w. It is easily seen that u_{2n} and v_{2n} are palindromes, and that $u_{2n+1} = \tilde{v}_{2n+1}$, where $\tilde{w}$ is the

reversal of w. The morphism μ can be extended to infinite words; it has two fixed points

$$t = abbabaabbaababbabaab \ldots = \mu(t),$$

$$E(t) = baababbaabbabaababba \ldots = \mu(E(t)).$$

The *Thue–Morse sequence* is the word t. Since u_n (respectively v_n) are the prefixes of length 2^n of t (respectively of $E(t)$), it is equivalent to say that t is the *limit* of the sequence $(u_n)_{n\geq 0}$ (for the usual topology on finite and infinite words), obtained by iterating the morphism μ.

There exist several other characterizations of the Thue–Morse sequence that can be found e.g. in Lothaire [8] and Salomaa [12].

More generally, we say that a word x is a *morphic* infinite word (with generator h) if there exists an integer $m \geq 1$ and a letter a such that

$$x = \lim_{n\to\infty} (h^{mn}(a)).$$

Of course, this implies that $x = h^m(x)$. Observe that the integer m is always bounded by the size of the alphabet.

A word w is *overlap-free* if it has no factor of the form $xuxux$ for some word u and some nonempty word x. Thue proved

Theorem 2.1 [15, Satz 5]. *The infinite word t is overlap-free.*

A morphism h is called *overlap-free* if $h(x)$ is overlap-free for every overlap-free word x. Clearly, the composition of two overlap-free morphisms is again overlap-free. Besides the two trivial morphisms, namely the identity and the morphism E that exchanges a and b, there is basically only one overlap-free morphism. Indeed, one has

Theorem 2.2 [15, Satz 15]. *Any overlap-free morphism h is of the form $h = \mu^k$ or $h = E \circ \mu^k$ for some integer $k \geq 0$.*

3. Results

The main result of this note is the following.

Theorem 3.1. *Let h be a morphism such that the word $h(abbabaab)$ is overlap-free. Then there exists an integer $k \geq 0$ such that $h = E \circ \mu^k$ or $h = \mu^k$.*

Thue's characterization of overlap-free morphisms is an immediate consequence of this theorem. We also get the following corollary.

Corollary 3.2. *A morphism h is overlap-free iff the word $h(abbabaab)$ is overlap-free.*

This corollary is of interest mainly when the morphism h is not explicitly known by its images $h(a)$ and $h(b)$, but when the word $h(abbabaab)$ can be computed. The theorem is an immediate consequence of the following proposition.

Proposition 3.3. *Let h be a nonerasing morphism such that $h(x)$ is overlap-free for any overlap-free word of length 3. Then there exists an integer $k \geq 0$ such that $h = \mu^k$ or $h = E \circ \mu^k$.*

Thus, in order to test whether a morphism h is overlap-free, it suffices either to consider the word $h(abbabaab)$ or to check that h is nonerasing and to consider the six words $h(aab)$, $h(aba)$, $h(abb)$, $h(baa)$, $h(bab)$ and $h(bba)$. If these are overlap-free, then h is of the indicated form and therefore is an overlap-free morphism. This statement is similar to a result by Karhumäki [6] saying that a binary morphism is cube-free iff it preserves cube-free words of length at most 10, and to a result by Crochemore [3] according to which a morphism over a three-letter alphabet is square-free iff it preserves square-free words of length at most 5.

Another consequence of the theorem that admits a short proof is the following result, first proved by Séébold [13]:

Theorem 3.4. *There are only two morphic infinite overlap-free words over a two-letter alphabet, namely the Thue–Morse sequence $\mathbf{t}$ and its opposite $E(\mathbf{t})$.*

Another way to state this is

Theorem 3.5. *Let $\mathbf{x}$ be an infinite overlap-free word that is a fixed point of a morphism h that is not the identity. Then $\mathbf{x}$ is the Thue–Morse sequence or its opposite.*

4. Proofs

We shall use the following lemmas, which are well known. The first two are due to Thue, and only the first has a slightly involved proof (see e.g. Lothaire [8]). The last lemma has been established independently by many people.

Lemma 4.1. *A word x is overlap-free iff $\mu(x)$ is overlap-free.*

Lemma 4.2. *If $x = cddyc'c'd'$ is an overlap-free word, where c, d, c', d' are letters and y is a word, then $dyc' \in \{ab, ba\}^*$.*

Lemma 4.3. *If x is an overlap-free word of length at least 5, then x contains a factor aa or bb.*

Lemma 4.4. *If x is an overlap-free word, then there exists an overlap-free word y and two words $u, v \in \{\varepsilon, a, b, aa, bb\}$ such that $x = u\mu(y)v$.*

Proof of Theorem 3.1. Assume that $w = h(abbabaab)$ is overlap-free. If $h(a) = \varepsilon$ then $h(bbab) = h(b)^3$ and w contains a cube. Thus $h(a)$ is nonempty and similarly $h(b)$ is nonempty. Consequently, h is nonerasing, and since *abbabaab* contains all overlap-free words of length 3 as factors, the morphism h fulfils the conditions of Proposition 3.3. □

Proof of Proposition 3.3. Set $h(a) = u$, $h(b) = v$. Observe that u and v do not start with the same letter, since otherwise $h(aab)$ contains an overlap. Similarly, u and v do not end with the same letter. Moreover, u and v do not start or end with *aa* nor with *bb*. Assume indeed for example that u starts with *aa*. Since either u or v ends with the letter a, one of the words uu or vu contains the cube a^3. Thus, if $|u| \geq 2$, it starts and ends with *ab* or *ba*, and the same holds for v.

The result holds if $|u| = |v| = 1$. Thus we assume that $|u| > 1$ or $|v| > 1$. We show first that then $|u| > 1$ and $|v| > 1$ and that both u and v are of even length. Assume indeed that $|u| > 1$. Then $|v| > 1$ since otherwise v is a single letter, say $v = a$, and then u starts with *ba* and ends with *ab*. Thus $h(aba)$ contains the factor *ababa*.

We now show that $|u|$ is even. If $|u| = 3$, then by the preceding discussion, one has $u = aba$ or $u = bab$. In the first case (the second is handled similarly), the word v starts and ends with the letter b. But then again $h(bab)$ contains the overlap *babab*. Next, assume $|u| \geq 5$. Then u has the form $u = pbaabs$ or $u = pabbas$ for some words p, s. Indeed, by Lemma 4.3, the word u contains a factor *aa* or *bb*, and we have seen that this factor is neither a prefix nor a suffix of u. In the first case (the second case is similar), the factor *baabspbaab* of the word $uu = pbaabspbaabs$ fulfils the conditions of Lemma 4.2, showing that *sp* has even length, and that $|u|$ is even.

We now prove that u and v are in $\{ab, ba\}^*$. If $|u| = 2$, then $u = ab$ or $u = ba$. If $|u| = 4$, then u starts and ends with *ab* or *ba*, and u is in $\{ab, ba\}^*$. If $|u| \geq 6$, then u contains a factor dd (with $d \in \{a, b\}$) which is neither a prefix nor a suffix. Further, we may assume that u starts with the letter a. We consider two cases, according to u ends with the letter a or b.

(i) If $u = awa$, then u admits a factorization $u = axddya$ for some letter d and some words x, y. Thus

$$uu = axddyaaxddya.$$

By Lemma 4.2, the words *dya* and *axd* are in $\{ab, ba\}^*$. Thus, $u \in \{ab, ba\}^*$.

(ii) If $u = awb$, then v starts and ends with *ba* (because $|v|$ is even, hence $|v| \geq 2$). Thus the word vuv contains the factor

$$baawbba.$$

By Lemma 4.2, the word u is in $\{ab, ba\}^*$. The proof that $v \in \{ab, ba\}^*$ is similar.

It follows that $u = \mu(u'), v = \mu(v')$ for some nonempty words u', v'. In view of Lemma 4.1, the words u' and v' are overlap-free. Define a morphism h' by $h'(a) = u', h'(b) = v'$. Then $h = \mu \circ h'$ and h' is nonerasing. Again by Lemma 4.1, the word $w' = h'(w)$ is overlap-free for every overlap-free word w of length 3 because $\mu(w') = h(w)$ is overlap-free. The result follows by induction on $|u| + |v|$. □

Proof of Theorem 3.4. This will be shown to be a consequence of Theorem 3.5. Indeed, if x is a morphic infinite word with generator h, then h is nontrivial and x is a fixed point of h^m, with $m = 1$ or $m = 2$. Thus $h^m = \mu^k$ or $h^m = E \circ \mu^k$ for some $k \geq 1$. If $m = 1$, the result is proved. Thus assume $h^2 = \mu^k$ or $h^2 = E \circ \mu^k$. Set

$$\alpha = |h(a)|_a, \qquad \beta = |h(a)|_b, \qquad \alpha' = |h(b)|_a, \qquad \beta' = |h(b)|_b.$$

Since the numbers $|h^2(a)|_a, |h^2(a)|_b, |h^2(b)|_a, |h^2(b)|_b$ are all equal to 2^{k-1}, it follows that

$$\alpha^2 + \alpha'\beta = \beta(\alpha + \beta') = \alpha'\beta + \beta'^2 = \alpha'(\alpha + \beta') = 2^{k-1}$$

which shows that the α and β all are equal, and that

$$\alpha^2 = \alpha'^2 = \beta^2 = \beta'^2 = 2^{k-2}.$$

Thus k is even. Set $k = 2l$. Then $|h(a)| = |h(b)| = 2^l$. Thus $h(a)$ and $h(b)$ are the prefixes of length 2^l of $\mu^k(a)$ and of $\mu^k(b)$ or vice versa, and consequently

$$h(a) = \mu^l(a), \qquad h(b) = \mu^l(b)$$

or vice versa. □

Proof of Theorem 3.5. Let $x = h(x)$ be an overlap-free word. By iterated application of Lemma 4.4, it is easily seen that x contains the factor *abbabaab*. Thus $h(abbabaab)$ is a factor of x and therefore is overlap-free. Thus, by Theorem 3.1, h is an overlap-free morphism, and $h = \mu^k$ or $h = E \circ \mu^k$ for some integer $k \geq 0$. The second case is ruled out by the fact that this morphism has no fixed point. Thus $h = \mu^k$, and since $k > 0$, one has $x = t$ or $x = E(t)$. □

Acknowledgement

We thank Antonio Restivo for having suggested an improvement of a previous version of Theorem 3.1 that leads to the present formulation, and the referees for their valuable comments.

References

[1] S. Arshon, Démonstration de l'existence des suites asymétriques infinies, Mat. Sb. 44 (1937) 769–777.

[2] J. Cassaigne, Counting overlap-free binary words, in: Enjalbert, Finkel and Wagner, eds., STACS '93, Lecture Notes in Computer Science 665 (Springer, Berlin, 1993) 216–225.

[3] M. Crochemore, Sharp characterizations of square-free morphisms, Theoret. Comput. Sci. 18 (1982) 221–226.
[4] E.D. Fife, Binary sequences which contain no *BBb*, Trans. Amer. Math. Soc. 261 (1980) 115–136.
[5] W.H. Gottschalk and G.A. Hedlund, A characterization of the Morse minimal set, Proc. Amer. Math. Soc. 15 (1964) 70–74.
[6] J. Karhumäki, On cube-free ω-words generated by binary morphisms, Discrete Appl. Math. 5 (1983) 279–297.
[7] Y. Kobayashi, Enumeration of irreducible binary words, Discrete Appl. Math. 20 (1988) 221–232.
[8] M. Lothaire, Combinatorics on Words (Addison-Wesley, Reading, MA, 1983).
[9] M. Morse, Recurrent geodesics on a surface of negative curvature, Trans. Amer. Math. Soc. 22 (1921) 84–100.
[10] M. Morse and G. Hedlund, Unending chess, symbolic dynamics and a problem in semigroups, Duke Math. J. 11 (1944) 1–7.
[11] A. Restivo and S. Salemi, Overlap-free words on two symbols, in: Nivat and Perrin, eds., Automata on Infinite Words, Lecture Notes on Computer Science 192 (Springer, Berlin, 1984) 198–206.
[12] A. Salomaa, Jewels of Formal Language Theory (Computer Science Press, Rockville, MD, 1981).
[13] P. Séébold, Sequences generated by infinitely iterated morphisms, Discrete Appl. Math. 11 (1985) 255–264.
[14] R. Shelton and R. Soni, Chains and fixing blocks in irreducible sequences, Discrete Math. 54 (1985) 93–99.
[15] A. Thue, Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen, Kra. Vidensk. Selsk. Skrifter. I. Mat.-Nat. Kl. Christiania 10 (1912) 1–67.